

# Mapping the Polarity of Tourist Opinions on Indonesian Destinations through Google Maps Reviews Using Supervised Learning Methods

Siti Miftahus Sa'adah <sup>1\*</sup>, Khothibul Umam <sup>2\*</sup>, Maya Rini Handayani <sup>3\*</sup>, Mokhammad Iklil Mustofa <sup>4\*</sup>

<sup>\*</sup> Teknologi Informasi, Fakultas Sains dan Teknologi, Universitas Islam Negeri Walisongo, Semarang, 50185, Indonesia  
[2208096099@student.walisongo.ac.id](mailto:2208096099@student.walisongo.ac.id) <sup>1</sup>, [khothibul\\_umam@walisongo.ac.id](mailto:khothibul_umam@walisongo.ac.id) <sup>2</sup>, [maya@walisongo.ac.id](mailto:maya@walisongo.ac.id) <sup>3</sup>,  
[iklil@walisongo.ac.id](mailto:iklil@walisongo.ac.id) <sup>4</sup>

## Article Info

### Article history:

Received 2025-06-16

Revised 2025-09-08

Accepted 2025-09-19

### Keyword:

*Tourist Opinion Polarity,  
Google Maps Reviews,  
Indonesian Tourist  
Destinations, Supervised  
Learning, Sentiment  
Classification, Text Mining,  
Natural Language Processing.*

## ABSTRACT

The advancement of information technology has transformed how individuals seek information and plan their travels, notably through online reviews of tourist attractions on platforms like Google Maps. However, these reviews do not always align with visitors' expectations, necessitating further analysis to comprehend the underlying sentiments. The objective of this research is to inspect the performance of multiple machine learning algorithms in executing sentiment analysis on user

generated reviews related to tourist attractions in Indonesia. The algorithms examined include Multinomial Naïve Bayes, Random Forest Classifier, Logistic Regression, Support Vector Machine, K-Nearest Neighbors, and Extra Trees Classifier. The research process encompasses data collection and labeling, data preprocessing, exploratory data analysis (EDA), Word Cloud visualization, feature extraction, classification implementation, and performance evaluation. Experimental results indicate that the K-Nearest Neighbors (KNN) algorithm attain the most accuracy and F1-score of 97%, indicating its effectiveness in categorizing text-based sentiment reviews sourced from the Google Maps platform.



This is an open access article under the [CC-BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.

## I. PENDAHULUAN

Efek positif yang signifikan telah dirasakan dampaknya karena kehadiran teknologi informasi terhadap perubahan pola perilaku Masyarakat dalam mengakses informasi dan merencanakan aktivitas pariwisata. Salah satu platform yang banyak dimanfaatkan untuk memperoleh informasi wisata adalah Google Maps. Melalui platform ini, pengguna dapat mengakses berbagai data seperti lokasi, gambar, serta ulasan dari pengunjung. Ulasan tersebut menjadi sumber referensi penting bagi wisatawan dalam memilih destinasi yang akan dikunjungi.

Konten ulasan dan penilaian yang diberikan oleh pengguna sangat beragam, mencakup ulasan yang bersifat positif, negatif, maupun netral. Untuk mengidentifikasi kecenderungan sentimen dalam ulasan pelanggan, apakah bersifat positif atau negatif, diperlukan analisis terhadap dataset yang tersedia. [1]. Walaupun ulasan dapat dijadikan sebagai salah satu acuan dalam memutuskan untuk

berkunjung, tidak jarang pengunjung mengalami ketidakpuasan karena kenyataan yang ditemui tidak sesuai dengan ekspektasi yang dibangun sebelumnya. Temuan ini sejalan dengan penelitian Khafidatul Ilmiyah dalam studinya yang berjudul "Pengaruh Ulasan Produk, Kemudahan, Kepercayaan, dan Harga terhadap Keputusan Pembelian pada Marketplace Shopee di Mojokerto," yang menyimpulkan bahwa tidak klesuruhan ulasan produk memberika hasil positif atas keputusan pembelian di platform Shopee. Ditemukan pula bahwa beberapa pengguna memberikan ulasan yang tidak sepenuhnya sesuai dengan penilaian sebenarnya sehingga dibutuhkan upaya penyaringan dan pengklasifikasi ulasan untuk memperoleh informasi yang lebih akurat[2].

Analisis sentimen menjadi teknik yang diimplementasikan oleh peneliti dengan sistematis mengidentifikasi, mengekstraksi, serta mengklasifikasikan opini atau ekspresi subjektif dalam teks terhadap suatu entitas tertentu, seperti produk, layanan, ataupun peristiwa. Analisis ini dilakukan

dengan tujuan utama menentukan kecenderungan sentimen yang muncul, apakah bersifat positif, negatif, maupun netral[3]. Analisis sentimen dapat didefinisikan sebagai proses untuk mengidentifikasi kandungan emosi dalam teks, baik berupa frasa, kalimat, paragraf, dokumen, maupun bentuk teks lainnya yang berasal dari suatu basis data atau dataset[4].

Dalam analisis sentimen, diperlukan penerapan algoritma data mining untuk mengklasifikasikan dataset ke dalam beberapa kategori label yang tersedia. Pada penelitian ini, penulis menerapkan beberapa metode algoritma, antara lain Naive Bayes, Logistic Regression, Random Forest, SVM, KNN, dan Extra Trees [5].

Berdasarkan uraian sebelumnya, penelitian ini ditujukan untuk mengukur tingkat akurasi dari berbagai metode yang digunakan serta untuk memperoleh model dengan akurasi terbaik dalam melaksanakan analisis sentimen terhadap ulasan sejumlah objek wisata di Indonesia melalui platform Google Maps.

Meskipun banyak studi telah membahas analisis sentimen dalam berbagai konteks, riset yang secara spesifik mengkaji ulasan destinasi wisata Indonesia melalui Google Maps masih sangat terbatas. Padahal, ulasan di platform ini memiliki potensi besar sebagai sumber insight terhadap pengalaman wisatawan secara real-time dan berbasis lokasi.

Penelitian ini memberikan kontribusi ilmiah dengan mengevaluasi performa enam algoritma supervised learning dalam mengklasifikasi sentimen wisata berbasis teks ulasan Google Maps berbahasa Indonesia. Strategi pelabelan berdasarkan skor rating serta eliminasi skor netral juga menjadi pendekatan praktis yang dapat meningkatkan kestabilan akurasi model. Dengan demikian, penelitian ini memperluas pemanfaatan metode NLP dan machine learning dalam sektor pariwisata berbasis data pengguna lokal.

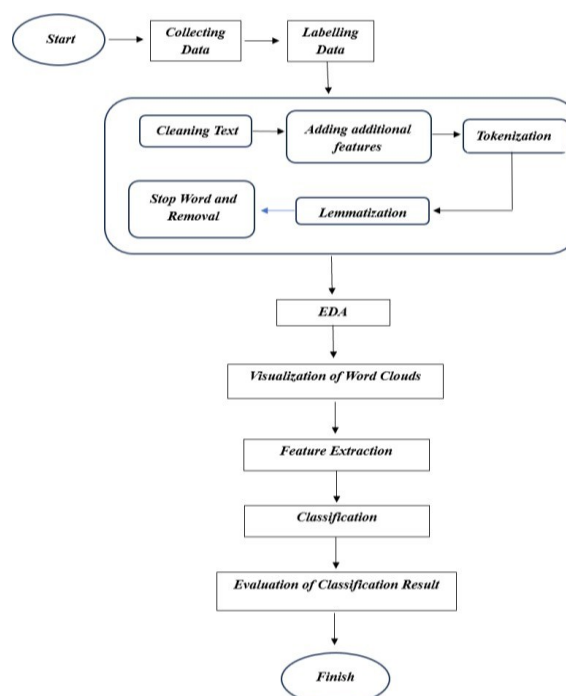
Selain itu, eksplorasi tuning parameter (hyperparameter tuning) juga dilakukan terhadap masing-masing algoritma supervised learning guna memperoleh akurasi maksimal. Eksplorasi kombinasi parameter ini menggunakan pendekatan Grid Search dan/atau Random Search, tergantung pada kompleksitas parameter setiap algoritma. Parameter yang dituning mencakup smoothing value ( $\alpha$ ) pada Naïve Bayes, kernel, C, dan gamma pada SVM, jumlah tetangga (k) pada KNN, serta jumlah pohon ( $n_{\text{estimators}}$ ) dan kedalaman maksimum ( $\text{max\_depth}$ ) pada Random Forest dan Extra Trees Classifier.

Oleh karena itu, penelitian ini memiliki novelty dalam hal (1) penggunaan Google Maps sebagai sumber utama data ulasan destinasi wisata Indonesia, (2) evaluasi kinerja berbagai algoritma klasifikasi modern pada dataset bahasa Indonesia, dan (3) strategi labeling berbasis skor rating sebagai pendekatan semi-terotomatis dalam ekstraksi sentimen wisata.

## II. METODE

Penelitian ini menggunakan Google Collaboratory sebagai platform untuk mendukung kegiatan pengolahan data dan proses klasifikasi. Beragam algoritma pembelajaran mesin yang diimplementasikan di penelitian ini antara lain Naïve Bayes, Random Forest, Regresi Logistik, SVM, KNN, serta Extra Trees Classifier. Pemilihan enam algoritma supervised learning dalam penelitian ini—yaitu Multinomial Naïve Bayes, Random Forest Classifier, Logistic Regression, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), dan Extra Trees Classifier—didasarkan pada karakteristik data teks ulasan yang bersifat tidak terstruktur dan beragam. Algoritma Naïve Bayes dipilih karena efisien dan cepat dalam mengklasifikasikan data teks dengan distribusi kata yang sederhana. Logistic Regression mampu memberikan hasil akurat dalam klasifikasi biner. Random Forest dan Extra Trees digunakan karena kemampuannya mengatasi overfitting dan menangani variabel penting secara otomatis. SVM dipilih karena efektif dalam menangani data berdimensi tinggi dan bekerja baik meskipun ukuran dataset tidak besar. Sedangkan KNN sesuai digunakan karena pendekatannya berbasis kemiripan antar teks, yang cocok untuk mendeteksi pola sentimen dalam ulasan.

Metode-metode tersebut tepat diterapkan pada data ini karena ulasan Google Maps bersifat teks pendek hingga menengah, dengan distribusi kata yang tidak seragam dan sentimen yang dapat diwakili oleh pemilihan kata-kata tertentu. Selain itu, data telah diproses melalui TF-IDF untuk mendukung representasi vektor yang optimal bagi algoritma-algoritma tersebut. Evaluasi komparatif dari masing-masing metode juga memberikan gambaran menyeluruh terhadap performa klasifikasi dalam konteks data ulasan wisata berbasis Bahasa Indonesia.



Gambar 1. Alur Penelitian

Gambar 1 mengungkapkan mengenai skema penelitian. Penelitian dilaksanakan melalui sejumlah tahapan utama, yaitu tahap akuisisi data, pemberian label pada data, serta tahap pra-pemrosesan data, analisis eksplorasi data (EDA), visualisasi word cloud, ekstraksi fitur, serta proses klasifikasi menggunakan berbagai algoritma, diikuti dengan evaluasi hasil klasifikasi. Pada tahap pra-pemrosesan data, dilakukan pemecahan prosedur menjadi beberapa langkah rinci, meliputi pembersihan data, penambahan fitur tambahan seperti panjang teks dan persentase tanda baca, proses tokenisasi, lemmatisasi, serta penghapusan kata-kata penghenti. Perincian tahapan ini dimaksudkan untuk memastikan tingkat ketelitian dan akurasi yang lebih tinggi dalam proses pengolahan dan analisis data, sehingga mendukung tercapainya hasil klasifikasi yang andal dan valid.

#### A. Data Collection

Data dalam penelitian ini diperoleh dari review dan rating pengguna di platform Google Maps, yang dikumpulkan melalui teknik web scraping[5]. Sebanyak 2.845 data telah dikumpulkan, yang mencakup rangkuman komentar dan penilaian yang paling relevan. Pengambilan data dilakukan menggunakan Google SERP API, dengan memanfaatkan Google Collaboratory sebagai platform untuk menjalankan proses pengumpulan data tersebut.

#### B. Data Labelling

Pelabelan data adalah proses menetapkan kecenderungan sentimen terhadap komentar yang diperoleh melalui teknik web scraping, guna mempersiapkan data untuk tahap analisis lebih lanjut[6]. Dalam penelitian ini, proses pelabelan dilakukan berdasarkan skor penilaian yang diberikan oleh pengguna terhadap ulasan destinasi wisata di Indonesia melalui platform Google Maps. Ulasan rentang skor 1 atau 2 diklasifikasikan dalam sentimen negatif dan diberi label 0, sedangkan skor 4 atau 5 digolongkan dalam sentimen positif dan diberi label 1. Sedangkan ulasan dengan skor 3 tidak dilibatkan dalam proses pelabelan, karena data tersebut menunjukkan ketidakstabilan dalam membangun model klasifikasi yang akurat.

#### C. Preprocessing

Pada penelitian ini, tahap pra-pemrosesan bertujuan untuk memproses, memperbaiki, dan menyaring data mentah secara cermat, sehingga data yang dihasilkan siap untuk digunakan dalam tahap analisis selanjutnya. Tahapan preprocessing terbagi menjadi lima proses yang mencakup:

1. Pembersihan Data: Tahapan pertama di proses pra-pemrosesan ialah pembersihan data (data cleaning) [6]. Proses dimulai dengan pembersihan data komentar hasil scraping, yaitu dengan menghapus berbagai karakter yang tidak

dibutuhkan, semisal emoji, nomor, tanda baca, serta karakter spesial lainnya. Selain itu, semua huruf kapital juga diubah menjadi huruf kecil[7].

2. Penambahan Fitur: Tahapan ini bertujuan untuk menganalisis frekuensi kemunculan tanda baca dalam teks ulasan serta menghitung persentase penggunaannya [8].
3. Tokenisasi: proses untuk memisahkan kalimat panjang menjadi komponen-komponen yang lebih kecil, seperti kata atau token, dengan cara yang terstruktur [8].
4. Lematisasi: Lematisasi merupakan proses normalisasi kata dengan mengubah berbagai bentuk kata turunan menjadi bentuk dasar (lemma) yang sesuai dengan konteksnya.
5. Penghapusan Stopwords: Tahapan akhir dalam pra-proses data adalah menyingkirkan macam kata yang dipandang tidak mengandung arti esensial, bersifat non-standar, atau tidak berdampak substansial terhadap hasil analisis[8].

#### D. Exploratory Data Analysis

Exploratory Data Analysis (EDA) merupakan langkah pertama pada pelaksanaan pengolahan data yang digunakan supaya meninjau dan mengetahui pola serta ciri khas data melalui pendekatan deskriptif dan visualisasi. Pada tahap ini, data hasil penelitian akan disusun dalam sebuah kerangka data (data frame) dan diklasifikasikan berdasarkan kategorinya, misalnya rating 1 hingga rating 5 akan dikelompokkan ke dalam kelas masing-masing. Setelah proses pengelompokan selesai, langkah berikutnya adalah melakukan identifikasi terhadap kemungkinan adanya data yang hilang, yang kemudian divisualisasikan dalam bentuk diagram batang untuk memudahkan interpretasi[9].

#### E. Visualisasi Word Clouds

Word cloud adalah bentuk visualisasi dari data teks yang merepresentasikan sekumpulan kata berdasarkan frekuensi kemunculannya, di mana ukuran masing-masing kata menunjukkan frekuensi kemunculannya dalam dokumen[10]. Frekuensi kemunculan kata berbanding lurus dengan ukurannya dalam visualisasi; kata dengan frekuensi tinggi dalam segi kemunculan disajikan dengan dimensi lebih besar. Word Clouds dibagi menjadi dua bagian di penelitian ini, yaitu Word Clouds untuk komentar positif dan Word Clouds untuk komentar negatif.

#### F. Feature Extraction

Feature extraction merupakan tahap reduksi dimensi yang bertujuan untuk mengubah data mentah menjadi representasi baru dengan jumlah variabel yang lebih sedikit. Dalam

penelitian ini, proses ekstraksi fitur dilakukan menggunakan metode vectorizer berbasis TF-IDF. Term Frequency (TF) merupakan metrik yang digunakan supaya menjumlahkan level kemunculan sebuah kata di dokumen teks; semakin tinggi frekuensinya, maka semakin besar pula nilai TF yang diperoleh.

Teknik ini terdiri dari dua komponen utama (1) :

1. Term Frequency (TF) mengukur seberapa sering suatu kata  $t$  muncul dalam dokumen  $d$ .

$$W(t,d) = TF(t,d) \quad (1)$$

Dimana  $TF(t,d)$  adalah frekuensi kemunculan kata  $t$  dalam dokumen  $d$

2. Inverse Document Frequency (IDF) mengukur seberapa unik atau jarang suatu kata muncul di seluruh dokumen.

$$IDF(t) = \log(N/df(t)) \quad (2)$$

Dimana  $N$  adalah jumlah total dokumen, dan  $df(t)$  adalah jumlah dokumen yang mengandung kata  $t$ .

Nilai TF-IDF diperoleh dengan mengalikan kedua komponen tersebut:

$$TF - IDF = TF(t,d) \times IDF(t) \quad (3)$$

Dengan pendekatan ini, kata-kata yang sering muncul dalam dokumen tertentu namun jarang muncul di dokumen lain akan memperoleh bobot tinggi, mencerminkan tingkat kepentingannya dalam dokumen tersebut.

### G. Classification

Setelah mengekstraksi fitur, langkah selanjutnya adalah melakukan klasifikasi. Dalam penelitian ini, berbagai algoritma yang digunakan secara terpisah.

1. *Multinomial Naïve Bayes*: Teorema Bayes ialah landasan algoritma Naïve Bayes yang mengespekulasikan bahwa masing-masing fitur atau atribut dalam proses klasifikasi bersifat saling otonom. Pendekatan ini dinilai efisien dalam membangun model klasifikasi karena memiliki kecepatan komputasi yang tinggi serta kebutuhan memori yang relatif rendah. Formula matematis untuk algoritma Multinomial Naïve Bayes disajikan pada Persamaan (4). [11]

$$P(c|d) = \frac{\prod_{i=1}^n P(w_i | c)^{f_i}}{P(d)}$$

$P(X|c)$ : probabilitas fitur  $X$  diberikan kelas  $c$

$N_c$  : jumlah total fitur dari kelas  $c$

$N$  : jumlah total fitur

$T_i$  : bobot kata ke- $i$

$v_{i,j}$ : frekuensi fitur ke- $i$  dalam kelas  $c$

$\sum_{i=1}^{n} \sum_{j=1}^n t_{i,j}$ : total bobot semua kata dalam kelas  $c$

$\alpha$ : smoothing parameter biasanya digunakan dalam Laplace Smoothing.

2. *Random Forest Classifier*: Random Forest Classifier ialah salah satu algoritma machine learning berbasis metode ensemble, yang beroperasi dengan membentuk sekumpulan pohon keputusan secara paralel, kemudian mengonsolidasi hasil prediksi dari masing-masing pohon guna menghasilkan output akhir yang lebih akurat [12]. Dalam melakukan klasifikasi, Random Forest Classifier mampu mengatasi ketidakseimbangan data, mengurangi overfitting, dan memberikan tingkat akurasi yang baik.
3. *Logistic Regression*: Logistic Regression merupakan algoritma machine learning yang sering diaplikasikan dalam tugas analisis sentimen. Dalam konteks ini, Algoritma ini menggunakan pendekatan berbasis statistik untuk memperkirakan kemungkinan sebuah teks tergolong pada ke sentimen positif atau sebaliknya berdasarkan variabel-variabel input yang tersedia [13]. Sebagai algoritma pembelajaran mesin, Regresi Logistik dapat dimanfaatkan supaya mengkategorikan data menjadi dua kelas, yaitu sentimen negatif dan sentimen positif.
4. *Support Vector Machine (SVM)*: Support Vector Machine (SVM) ialah algoritma dalam machine learning yang secara luas diterapkan dalam menyelesaikan permasalahan klasifikasi dan regresi. Pada lingkup klasifikasi, SVM dikenal mampu memberikan performa yang optimal meskipun pada dataset berukuran kecil. Selain itu, algoritma ini juga efektif dalam mengatasi permasalahan overfitting melalui penerapan teknik regularisasi. Dalam penelitian ini digunakan varian non-linear dari algoritma SVM, yang mengandalkan fungsi kernel supaya merepresentasikan data ke dalam ruang fitur berukuran lebih tinggi, yang menjadikan pemisahan

antar kelas menjadi lebih optimal. Adapun jenis kernel yang digunakan dalam pemodelan klasifikasi ini adalah kernel linear[14]. Fungsi prediksi dalam SVM non-linear dengan kernel dapat dituliskan dalam persamaan (5) sebagai berikut:

$$y(x) = \sum(\alpha_i * y_i * K(x_i, x)) + b \quad (5)$$

Informasi:

$y(x)$ : fungsi prediksi

$\alpha_i$ : koefisien lagrange

$y_i$ : label kelas

$K$ : fungsi kernel  $x_i$ ,

$x$ : vektor fitur input

$b$ : bias

Rumus kernel linear dapat dilihat dalam persamaan (6).

$$K(x_i, x) = x_i^T * x \quad (6)$$

5. *K-Nearest Neighbor*: K-NN adalah metode pembelajaran mesin yang bersifat non-parametrik dan menerapkan pendekatan pembelajaran malas (lazy learning). Ketidakparametrian algoritma ini menunjukkan bahwa tidak ada asumsi tertentu mengenai distribusi data yang mendasari proses klasifikasi. Sedangkan, prinsip pembelajaran malas mengindikasikan bahwa K-NN tidak mengandalkan data latih untuk membangun model secara langsung. Pada proses klasifikasi, algoritma ini menetapkan nilai K sebagai total tetangga paling dekat yang menjadi acuan pada proses pengambilan keputusan. Setelah itu, rentang antara data baru dan data pelatihan dijumlahkan melalui metrik jarak tertentu yang telah ditentukan sebelumnya. K-NN kemudian memilih K tetangga terdekat dengan jarak terkecil, dan klasifikasi dilakukan berdasarkan mayoritas label kelas dari tetangga yang dipilih tersebut[13].
6. *Extra Trees Classifier*: Extra Trees Classifier adalah algoritma pembelajaran mesin yang termasuk dalam kategori metode ensemble, yang juga merupakan turunan dari algoritma Random Forest. Algoritma ini beroperasi dengan membentuk sejumlah pohon keputusan yang dibangun berdasarkan sampel data dan fitur yang dipilih secara acak, kemudian

melakukan pemungutan suara mayoritas untuk menghasilkan prediksi akhir[15].

7. *Evaluation of Classification Results*: Langkah terakhir dalam penelitian ini adalah mengevaluasi kinerja algoritma yang diterapkan dalam proses klasifikasi. Evaluasi bertujuan untuk menilai sejauh mana algoritma tersebut dapat memprediksi dataset dengan akurat. Teknik yang digunakan untuk evaluasi adalah Confusion Matrix, sebuah tabel yang sering digunakan dalam pembelajaran mesin untuk membandingkan hasil prediksi model dengan data aktual[14].

Tabel I  
CONFUSION MATRIX

Prediksi	Nilai Aktual	
	Positif	Negatif
Positif	TP	FN
Negatif	FP	TN

Ini adalah tabel matriks kebingungan yang memberikan informasi berikut:

TP (True Positive): Merujuk pada total data yang dikategorikan berupa positif dan memang valid termasuk dalam kategori positif berdasarkan kondisi sebenarnya.

TN (True Negative): Menunjukkan total data yang diperkirakan menjadi negatif dan selaras berdasarkan kondisi aktual yang juga negatif.

FP (False Positive): Menunjukkan kasus ketika data diprediksi sebagai positif, padahal dalam kenyataannya termasuk kategori negatif.

FN (False Negative): Merujuk terhadap total data yang diperkirakan sebagai negatif, namun sebetulnya merupakan data dengan kategori positif. Dengan memanfaatkan nilai TP, TN, FP, dan FN, peneliti dapat menghitung berbagai matriks evaluasi untuk menilai performa model, semisal akurasi, presisi, recall, dan F1-Score. Persamaan supaya menghitung setiap matriks dapat dilihat pada rumus (7), (8), (9), dan (10) sebagai berikut:

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN} \quad (7)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (8)$$

$$Recall = \frac{TP}{TP + FN} \quad (9)$$

$$F1 \text{ Score} = 2 \times \frac{Recall \times Precision}{Recall + Precision} \quad (10)$$

### III. HASIL DAN PEMBAHASAN

Bagian ini memaparkan seluruh hasil penelitian, yang mencakup berbagai tahapan mulai dari pengumpulan data, proses pelabelan data, praproses data, analisis data eksploratori (EDA), visualisasi word cloud, ekstraksi fitur, proses klasifikasi, hingga evaluasi terhadap hasil klasifikasi.

#### A. Data Collection

Data diperoleh melalui pemanggilan API, yang menghasilkan sebanyak 2.845 entri, dengan rincian field dan record disajikan sebagai berikut.

rating		Review	Unnamed :2	Unnamed :3
0	5	Kesini hr minggu jam 9 pagi dg tiket explorer ...	NaN	NaN
1	4	Tempat wisata dan edukasi yang cocok didatangi ...	NaN	NaN
2	4	Halooo.. Ini dia, taman safari prigen di daera ...	NaN	NaN
3	5	Datang kesini pas moment liburan sekolah dihar ...	NaN	NaN
4	5	Sdh ketiga kalinya kesini dgn keluarga.. Beli ...	NaN	NaN

Gambar 2. Koleksi data frame awal

Dari data yang diperoleh, peneliti memilih dua kolom utama, yaitu Rating dan Review. Kolom Review akan digunakan sebagai sumber untuk ekstraksi fitur, sementara kolom Rating akan dikonversi menjadi label pada tahap pemrosesan berikutnya.

#### B. Data Labelling

Setelah tahap pengambilan data selesai dilakukan, data yang telah terkumpul selanjutnya diproses pada tahap pelabelan guna mengidentifikasi apakah setiap komentar termasuk ke dalam sentimen positif atau negatif.

rating		Review	Unnamed :2	Unnamed :3
0	5	Kesini hr minggu jam 9 pagi dg tiket explorer ...	NaN	NaN
1	4	Tempat wisata dan edukasi yang cocok didatangi ...	NaN	NaN
2	4	Halooo.. Ini dia, taman safari prigen di daera ...	NaN	NaN
3	5	Datang kesini pas moment liburan sekolah dihar ...	NaN	NaN
4	5	Sdh ketiga kalinya kesini dgn keluarga.. Beli ...	NaN	NaN

Gambar 3. Hasil labelling data

Pada gambar 4 dalam penelitian ini, nilai 0 diberikan kepada rating 1 dan 2 yang menunjukkan label negatif, sedangkan nilai 1 diberikan kepada rating 4 dan 5 yang menunjukkan label positif, dan rating 3 kita tidak masukkan

karena saat melakukan modeling data akurasi tidak konsisten dan kurang akurat.

#### C. Preprocessing

##### 1. Pembersihan Data

Rating		Review	Unnnamed :2	Unnnamed :3	Cleaned_text	Label
0	5	Kesini hr minggu jam 9 pagi dg tiket explorer ...	NaN	NaN	Kesini hr minggu jam 9 pagi dg tiket explorer ...	1
1	4	Tempat wisata dan edukasi yang cocok didatangi ...	NaN	NaN	Tempat wisata dan edukasi yang cocok didatangi ...	1
2	4	Halooo.. Ini dia, taman safari prigen di daera ...	NaN	NaN	Halooo.. Ini dia, taman safari prigen di daera ...	1
3	5	Datang kesini pas moment liburan sekolah dihar ...	NaN	NaN	Datang kesini pas moment liburan sekolah dihar ...	1
4	5	Sdh ketiga kalinya kesini dgn keluarga.. Beli ...	NaN	NaN	Sdh ketiga kalinya kesini dgn keluarga.. Beli ...	1

Gambar 4. Hasil Pembersihan Data

Merujuk pada Gambar 5, teks pada kolom Review telah melalui proses pembersihan dan hasilnya disimpan dalam kolom baru bernama cleaned\_text. Pada tahap pelabelan, data dengan nilai Rating sebesar 3 dihapus karena dianggap netral dan tidak menunjukkan kecenderungan terhadap sentimen positif maupun negatif. Selanjutnya, proses pelabelan sentimen dilakukan dengan menetapkan nilai 1 untuk sentimen positif dan 0 untuk sentimen negatif, yang kemudian disimpan dalam kolom label.

##### 2. Penambahan Fitur

Rating		Review	Unnnamed :2	Unnnamed :3	Cleaned_text	Label	Review -len	punct
0	5	Kesini hr minggu jam 9 pagi dg tiket explorer ...	NaN	NaN	Kesini hr minggu jam 9 pagi dg tiket explorer ...	1	655	4.6
1	4	Tempat wisata dan edukasi yang cocok didatangi ...	NaN	NaN	Tempat wisata dan edukasi yang cocok didatangi ...	1	1461	2.8
2	4	Halooo.. Ini dia, taman safari prigen di daera ...	NaN	NaN	Halooo.. Ini dia, taman safari prigen di daera ...	1	2735	4.3
3	5	Datang kesini pas moment liburan sekolah dihar ...	NaN	NaN	Datang kesini pas moment liburan sekolah dihar ...	1	1039	1.9
4	5	Sdh ketiga kalinya kesini dgn keluarga.. Beli ...	NaN	NaN	Sdh ketiga kalinya kesini dgn keluarga.. Beli ...	1	630	2.4

Gambar 5. Hasil Penambahan Fitur

Pada proses ini, dilakukan penambahan fitur berupa panjang teks review (yang dimasukkan pada kolom "Review\_len") dan persentase kemunculan tanda baca (yang dimasukkan pada kolom "punc").

##### 3. Tokenisasi



Rating		Review	Unnamed :2	Unnamed :3	Cleaned_text	Label	Review -len	punct	tokens
0	5	Kesini hr minggu jam 9 pagi dg tiket explorer ...	NaN	NaN	Kesini hr minggu jam 9 pagi dg tiket explorer ...	1	655	4.6	[kesini, hr, minggu, jam, pagi, dg, tiket,exp ...]
1	4	Tempat wisata dan edukasi yang cocok didatangi ...	NaN	NaN	Tempat wisata dan edukasi yang cocok didatangi ...	1	1461	2.8	[tempat, wisata, dan, edukasi, yang, cocok,di, ...]
2	4	Halooo, Ini dia, taman safari prigen di daera ...	NaN	NaN	Halooo, Ini dia, taman safari prigen di daera ...	1	2735	4.3	[halooo, Ini dia, taman safari prigen di daera, ...]
3	5	Datang kesini pas moment liburan sekolah dihar ...	NaN	NaN	Datang kesini pas moment liburan sekolah dihar ...	1	1039	1.9	Datang, kesini, pas, moment ,liburan ,sekolah...
4	5	Sdh ketiga kalinya kesini dgn keluarga.. Beli ...	NaN	NaN	Sdh ketiga kalinya kesini dgn keluarga.. Beli ...	1	630	2.4	Sdh, ketiga, kalinya, kesini ,dgn, keluarga,...

Gambar 6. Hasil Tokenisasi

Pada tahap ini, teks diuraikan menjadi unit-unit kata yang disebut token. Adanya tahap ini bermaksud supaya medari proses tokenisasi adalah untuk memecahkan kata-kata dalam teks sehingga dapat difilter dan digunakan pada tahap pemrosesan berikutnya.

#### 4. Lematisasi dan Penghapusan Stopwords

tokens	lemmatized_review
[kesini, hr, minggu, jam, pagi, dg, tiket, exp...]	kesini hr minggu jam pagi dg tiket explorer id...
[tempat, wisata, dan, edukasi, yang, cocok, di...]	wisata edukasi cocok didatangi keluarga tempat...
[halooo, ini, dia, taman, safari, prigen, di, ...]	halooo taman safari prigen daerah pasuruan pik...
[datang, kesini, pas, moment, liburan, sekolah...]	kesini pa moment liburan sekolah dihari kerja ...
[sdh, ketiga, kalinya, kesini, dgn, keluarga, ...]	sdh ketiga kalinya kesini dgn keluarga beli ti...

Gambar 7. Hasil Lematisasi dan Penghapusan Stopwords

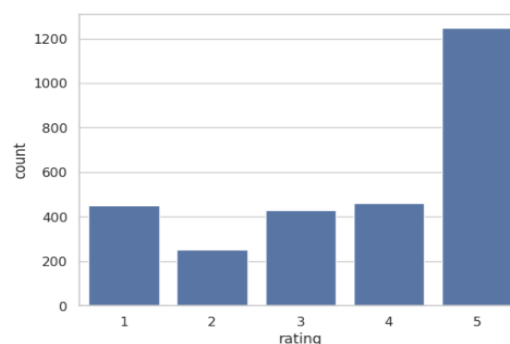
Pada bagian ini dilakukan langsung dilakukan 2 proses sekaligus yang hasilnya dimasukkan ke dalam kolom bernama “lemmatized\_review”.

#### D. Exploratory Data Analysis (EDA)

Pada tahap ini, data penelitian berhasil disusun dalam bentuk data frame dan diklasifikasikan ke dalam masing-masing kelas berdasarkan nilai rating, di mana rating 1 ditempatkan pada kelas 1, rating 2 pada kelas 2, rating 3 pada kelas 3, rating 4 pada kelas 4, dan rating 5 pada kelas 5.

Input data has 2844 rows and 2 columns  
 rating 1.0 = 451 rows  
 rating 2.0 = 254 rows  
 rating 3.0 = 431 rows  
 rating 4.0 = 461 rows  
 rating 5.0 = 1247 rows

Gambar 8. Detail data kelas



Gambar 9. Detail data kelas

Gambar 9 dan 10 menyajikan informasi bahwa dalam penelitian ini terdapat data input sebanyak 2844 baris dan 2 kolom, dengan peringkat tertinggi adalah rating 5 dengan jumlah baris sebanyak 1247, diikuti oleh rating 4 dengan jumlah baris sebanyak 461, rating 1 dengan jumlah baris 451, rating 3 sebanyak 431 dan terakhir rating 2 dengan jumlah baris sebanyak 254.

#### E. Visualisasi Word Clouds



Gambar 10. Word Clouds Positive Comments



Gambar 11. Word Clouds Negative Comments

Sebagaimana dijelaskan pada bagian Metode Penelitian, penelitian ini berhasil menghasilkan dua visualisasi word cloud, yaitu visualisasi dari ulasan bersentimen positif dan ulasan bersentimen negatif. Sesuai Gambar 6 dan Gambar 7, kata-kata dengan frekuensi kemunculan terbanyak pada word cloud untuk ulasan positif dan negatif di antaranya adalah 'dan', 'yang', 'ada', 'untuk', dan 'dari'. Kemunculan kata-kata tersebut mencerminkan frekuensi tinggi dalam teks.

#### F. Feature Extraction

review	len	punct	0	1	2	3	4	5	6	7	...	4157	4158	4159	4160	4061	4162	4163	4164	4165	4166
0	825	14	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	347	6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	252	7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	242	6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	156	7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Gambar 12. Hasil Ekstraksi Fitur

Gambar 13 menampilkan hasil ekstraksi fitur dari kolom `lemetized_review` menggunakan algoritma TF-IDF. Pada tahap ini, data yang sebelumnya berbentuk string atau teks dikonversi menjadi representasi vektor, yang menghasilkan peningkatan jumlah fitur yang dapat digunakan dalam proses analisis lebih lanjut.

#### G. Classification and Evaluation of Classification Results

Sebelum memasuki tahap klasifikasi, data penelitian dibagi menjadi dua subset, yakni data latih dan data uji, dengan proporsi 70:30. Artinya, 70% dari keseluruhan data digunakan untuk proses pelatihan model, sedangkan 30% lainnya dimanfaatkan guna pengujian kinerja model. Berdasarkan pembagian tersebut, diperoleh hasil klasifikasi sebagai berikut:

```
Jumlah fold yang dipakai: 7
Naive Bayes
accuracy - 0.9047619047619048
precision - 0.9523809523809523
recall - 0.9285714285714286
f1 - 0.9238095238095239
```

```
Logistic Regression
accuracy - 0.8809523809523808
precision - 0.8809523809523808
recall - 1.0
f1 - 0.9238095238095239
```

```
Random Forest
accuracy - 0.8571428571428571
precision - 0.9523809523809523
recall - 0.9285714285714286
f1 - 0.9238095238095239
```

```
SVM
accuracy - 0.9047619047619048
precision - 0.9523809523809523
recall - 0.9285714285714286
f1 - 0.9238095238095239
```

```
KNN
accuracy - 0.9523809523809523
precision - 0.9523809523809523
recall - 1.0
f1 - 0.9714285714285714
```

```
Extra Trees
accuracy - 0.9047619047619048
precision - 0.9523809523809523
recall - 0.9285714285714286
f1 - 0.9238095238095239
```

Gambar 13. Hasil Skor Model dari Beberapa Algoritma

Eksperimen yang dilakukan menggunakan beberapa algoritma supervised learning, yaitu Naïve Bayes, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Random Forest, Logistic Regression, dan Extra Trees Classifier. Setiap model dievaluasi menggunakan metrik akurasi, precision, recall, dan F1-score.

Hasil menunjukkan bahwa KNN menghasilkan nilai F1-score tertinggi, yang mengindikasikan performa seimbang dalam mengenali kelas positif, netral, dan negatif. Hal ini menunjukkan bahwa KNN dapat bekerja efektif pada dataset ulasan dengan distribusi yang tidak sepenuhnya seimbang. Namun demikian, KNN memiliki kekurangan dalam hal efisiensi karena memerlukan perhitungan jarak dengan semua titik data pada saat prediksi, sehingga kurang optimal untuk dataset berskala besar.

SVM dan Random Forest juga menunjukkan performa yang kompetitif, terutama setelah dilakukan hyperparameter tuning. SVM, dengan kernel RBF dan parameter yang diatur, mampu memisahkan kelas dengan margin optimal. Random Forest, sebagai model ansambel berbasis pohon keputusan, menawarkan keunggulan dalam menangani fitur yang saling bergantung dan bersifat non-linear.

Sementara itu, Naïve Bayes meskipun sederhana dan cepat, cenderung menghasilkan performa yang lebih rendah, diduga karena asumsi independensi antar fitur tidak sepenuhnya terpenuhi dalam data ulasan teks.

Logistic Regression menunjukkan kinerja yang cukup baik namun sedikit lebih rendah dibandingkan model non-linear. Hal ini dapat disebabkan oleh keterbatasannya dalam menangani relasi kompleks antar fitur yang umum ditemukan dalam representasi vektor teks.

Hasil eksperimen menunjukkan bahwa proses hyperparameter tuning mampu meningkatkan performa algoritma klasifikasi. Sebagai contoh, akurasi algoritma SVM meningkat dari 84,5% menjadi 87,2% setelah dilakukan tuning dengan penggunaan kernel RBF, parameter  $C=10$ , dan  $\gamma=0.01$ . Sementara itu, algoritma Random Forest juga mengalami peningkatan akurasi dari 82% menjadi 85,3% setelah penyetelan parameter  $n\_estimators=200$  dan  $max\_depth=25$ . Dengan demikian, proses tuning parameter terbukti berkontribusi dalam mengoptimalkan hasil klasifikasi sentimen terhadap ulasan wisata.

Dari hasil ini dapat disimpulkan bahwa pemilihan algoritma harus mempertimbangkan karakteristik data dan kebutuhan aplikasi. Selain itu, proses tuning parameter terbukti mampu meningkatkan performa model secara signifikan.

#### IV. KESIMPULAN

Kesimpulan yang dapat ditarik selaras dengan hasil penelitian ialah algoritma K-Nearest Neighbors (KNN) menghasilkan nilai F1-score tertinggi di antara seluruh



algoritma yang diuji yaitu sebesar 97% . Keunggulan ini turut didukung oleh nilai F1-score yang juga mencapai angka 97%, mengindikasikan konsistensi performa model dalam klasifikasi. Keterbatasan yang ditemukan di penelitian ini mencakup ketidakseimbangan rasio antara data berlabel negatif dan positif, yang dikenal dengan istilah oversampling. Hal ini dapat menimbulkan masalah seperti overfitting dan underfitting pada model[16].

#### DAFTAR PUSTAKA

- [1] E. H. Muktafin, K. Kusri, and E. T. Luthfi, "Analisis Sentimen pada Ulasan Pembelian Produk di Marketplace Shopee Menggunakan Pendekatan Natural Language Processing," *J. Eksplorasi Inform.*, vol. 10, no. 1, pp. 32–42, Sep. 2020, doi: 10.30864/eksplorasi.v10i1.390.
- [2] K. Ilmiyah and I. Krishnawati, "Pengaruh Ulasan Produk, Kemudahan, Kepercayaan, Dan Harga Terhadap Keputusan Pembelian Pada Marketplace Shopee Di Mojokerto," *Mak. J. Manaj.*, vol. 6, no. 1, pp. 31–42, Jun. 2020, doi: 10.37403/mjm.v6i1.143.
- [3] M. Wankhade, A. C. S. Rao, and C. Kulkarni, "A survey on sentiment analysis methods, applications, and challenges," *Artif. Intell. Rev.*, vol. 55, no. 7, pp. 5731–5780, Oct. 2022, doi: 10.1007/s10462-022-10144-1.
- [4] E. Kontopoulos, C. Berberidis, T. Dergiades, and N. Bassiliades, "Ontology-based sentiment analysis of twitter posts," *Expert Syst. Appl.*, vol. 40, no. 10, pp. 4065–4074, Aug. 2013, doi: 10.1016/j.eswa.2013.01.001.
- [5] F. F. Irfani, "Analisis Sentimen Review Aplikasi Ruangguru Menggunakan Algoritma Support Vector Machine," *JBMI (Jurnal Bisnis, Manajemen, dan Inform.)*, vol. 16, no. 3, pp. 258–266, Feb. 2020, doi: 10.26487/jbmi.v16i3.8607.
- [6] D. Efriadi, R. Rahmaddeni, A. Agustin, and J. Junadhi, "Prediksi Penambahan Piutang Iuran Jaminan Sosial Ketenagakerjaan menggunakan Algoritma K-Nearest Neighbor," *EduMatic J. Pendidik. Inform.*, vol. 6, no. 1, pp. 49–57, Jun. 2022, doi: 10.29408/edumatic.v6i1.5255.
- [7] P. Aditiya, U. Enri, and I. Maulana, "Analisis Sentimen Ulasan Pengguna Aplikasi Myim3 Pada Situs Google Play Menggunakan Support Vector Machine," *JURIKOM (Jurnal Ris. Komputer)*, vol. 9, no. 4, p. 1020, Aug. 2022, doi: 10.30865/jurikom.v9i4.4673.
- [8] M. F. A. dan K. M. Lhaksana, "Perbandingan Metode Decision Tree dan Support Vector Machine untuk Analisis Sentimen pada Instagram Mengenai Kinerja PSSI," *e-Proceeding Eng.*, vol. Volume 7, pp. 9936–9948, 2020.
- [9] D. Agustina dan F. Rahmah, "Analisis Sentimen pada Sosial Media Twitter terhadap MRT Jakarta Menggunakan Machine Learning," *Insearch Inf. Syst. Res. J.*, vol. Volume 2, 2022.
- [10] A. I. Tanggraeni dan M. N. N. Sitokdana, "Analisis Sentimen Aplikasi E-Government Pada Google Play Menggunakan Algoritma Naïve Bayes," vol. Volume 9, pp. 785–795, 2022.
- [11] N. L. P. M. Putu, Ahmad Zuli Amrullah, and Ismarmiaty, "Analisis Sentimen dan Pemodelan Topik Pariwisata Lombok Menggunakan Algoritma Naive Bayes dan Latent Dirichlet Allocation," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 5, no. 1, pp. 123–131, Feb. 2021, doi: 10.29207/resti.v5i1.2587.
- [12] A. N. S. Budiman, A. Sunyoto, "Analisa Performa Penggunaan Feature Selection untuk Mendeteksi Intrusion Detection Systems dengan Algoritma Random Forest Classifier," *Sist. J. Sist. Inf.*, vol. Vol. 10, N, pp. 752–760, 2021, [Online]. Available: <http://sistemasi.ftik.unisi.ac.id>
- [13] R. A. F. Fazrin, O. Nurul Pratiwi, "Perbandingan Algoritma K-Nearest Neighbor dan Logistic Regression pada Analisis Sentimen terhadap Vaksinasi Covid-19 pada Media Sosial Twitter dengan Pelabelan Vader dan Textblob," in *e-Proceeding of Engineering*, 2023, pp. 1596–1604. [Online]. Available: <https://openlibrarypublications.telkomuniversity.ac.id/index.php/engineering/article/view/19960>
- [14] M. F. Asshiddiqi dan K. M. Lhaksana, "Perbandingan Metode Decision Tree dan Support Vector Machine untuk Analisis Sentimen pada Instagram Mengenai Kinerja PSSI," in *e- Proceeding of Engineering*, 2020, pp. 9936–9948.
- [15] A. S. A. S. K. Dirjen, P. Riset, D. Pengembangan, R. Dikti, S. Khomsah, "Model Text-Preprocessing Komentar Youtube Dalam Bahasa Indonesia," vol. Vol. 1, No, pp. 648–654, 2017.
- [16] R. Tuntun, K. Kusri, and K. Kusnawi, "Analisis Perbandingan Kinerja Algoritma Klasifikasi dengan Menggunakan Metode K- Fold Cross Validation," *J. MEDIA Inform. BUDIDARMA*, vol. 6, no. 4, p. 2111, Oct. 2022, doi: 10.30865/mib.v6i4.4681.